Anomaly Attribution with Likelihood Compensation

Tsuyoshi ("Ide-san") Idé, Amit Dhurandhar, Jiří Navrátil, Moninder Singh, Naoki Abe

IBM T. J. Watson Research Center

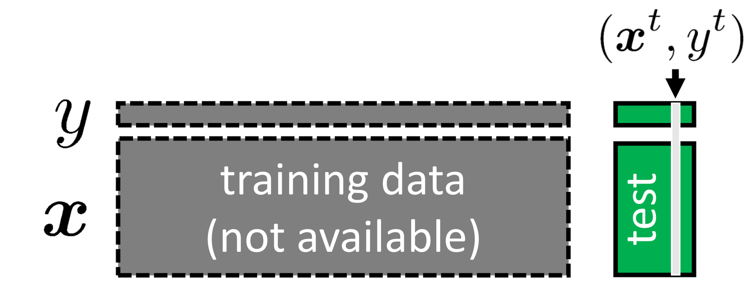
Slides available: https://ide-research.net

Technical task: anomaly attribution for black-box regression

- **Task**: Attribute deviation from black-box prediction $f(\boldsymbol{x})$ to each input variable
- **Background**: Most of XAI methods are designed to explain $f(\boldsymbol{x})$, not deviations
- **Solution**: New notion of "likelihood compensation"

Black-box regression func

real-valued $y = f(\boldsymbol{x})$ Multivariate vector in $\mathbb{R}^M$



Input

Test sample(s) showing anomaly / deviation

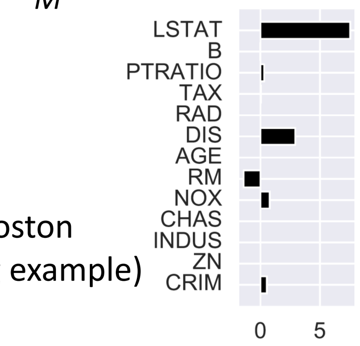
$\{ (\boldsymbol{x}^t, y^t) \}$

Black-box regression function $y = f(\boldsymbol{x})$

Likelihood compensation (LC) algorithm

Output

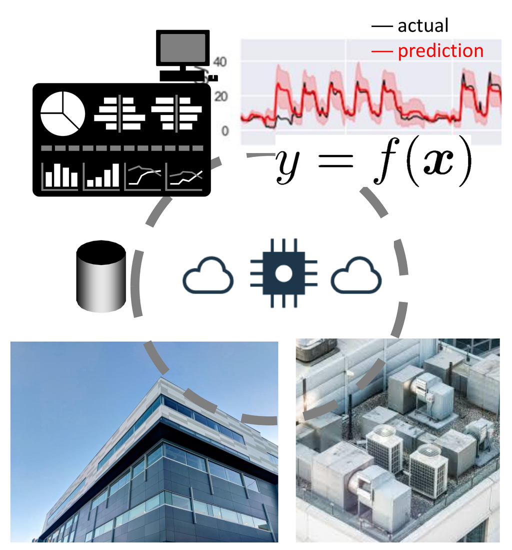
responsibility score computed locally at $(\boldsymbol{x}^t, y^t)$: $\delta_1, \ldots, \delta_M$



(from Boston Housing example)

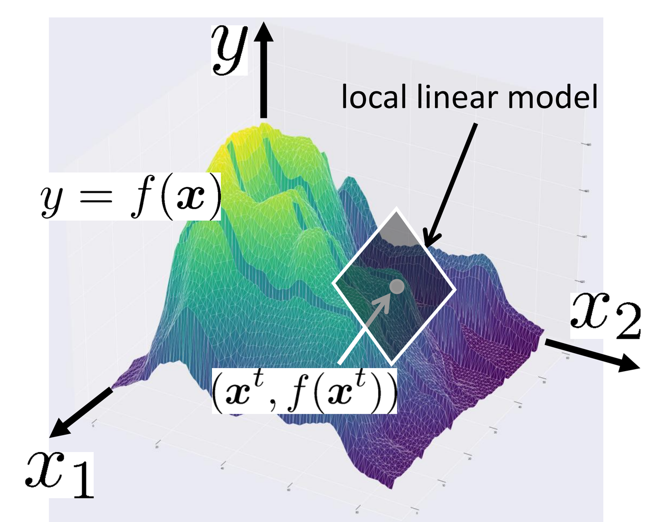
Use-case example: Building energy management

- y: building energy consumption
- **x**: Temperature, humidity, day of week, month, room occupancy, etc.
- Why black-box?
  - Multiple players (AI vendor / Sier / HVAC constructor)
  - Proprietary technologies

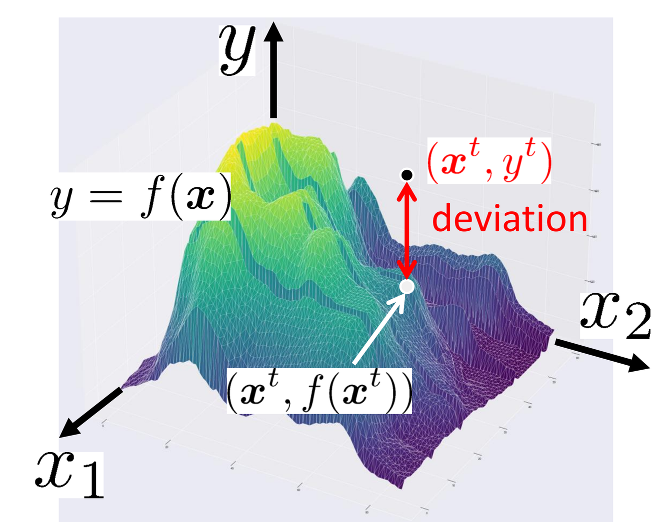


Local surrogate model for $f(\boldsymbol{x})$ alone cannot explain deviations. Defining responsibility score as "horizontal deviation"

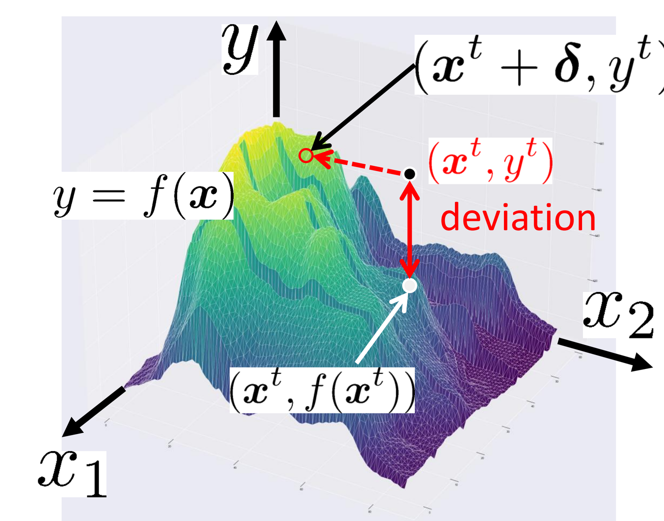
Local surrogate is to explain $f(\boldsymbol{x})$



Anomaly attribution is for $f(\boldsymbol{x})$ - y



δ "likelihood compensation"



LC as optimal perturbation to $\boldsymbol{x}^t$

- Likelihood compensation $\boldsymbol{\delta}$: A perturbation to $\boldsymbol{x}^t$ such that $\boldsymbol{x}^t + \boldsymbol{\delta}$ achieves the best possible fit to the model
  - LC seeks a best possible fit by correcting $\boldsymbol{x}^t$ under a certain regularization

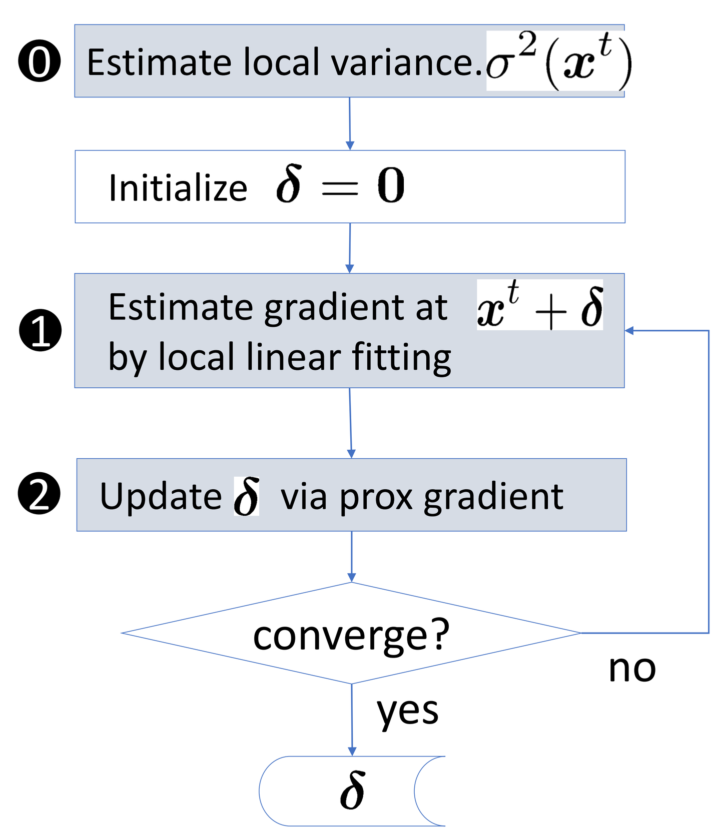
$$\boldsymbol{\delta} = \arg\max_{\boldsymbol{\delta}} \left[\ln \left\{ p(y^t \mid f(\boldsymbol{x}^t + \boldsymbol{\delta})) \; p(\boldsymbol{\delta}) \right\} \right],$$

Gaussian   elastic net

- The main optimization problem

$$\min_{\boldsymbol{\delta}} \left\{ \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2\sigma_t^2} + \frac{1}{2}\lambda \|\boldsymbol{\delta}\|_2^2 + \nu \|\boldsymbol{\delta}\|_1 \right\},$$

(λ and ν are constant)



0. Local variance estimation

- No extra test sample available → use constant variance (no choice!)

$$p(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid f(\boldsymbol{x}), \sigma^2)$$

- Some amount of test samples available → locally weighted maximum likelihood

$$\sigma^2(\boldsymbol{x}^t) = \max_{\sigma^2} \sum_{n=1}^{N_{\text{heldout}}} w_n(\boldsymbol{x}^t) \left\{ \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y^{(n)} - f(\boldsymbol{x}^{(n)}))^2}{2\sigma^2} \right\},$$

Gaussian kernel defined for the specific test sample $\boldsymbol{x}^t$

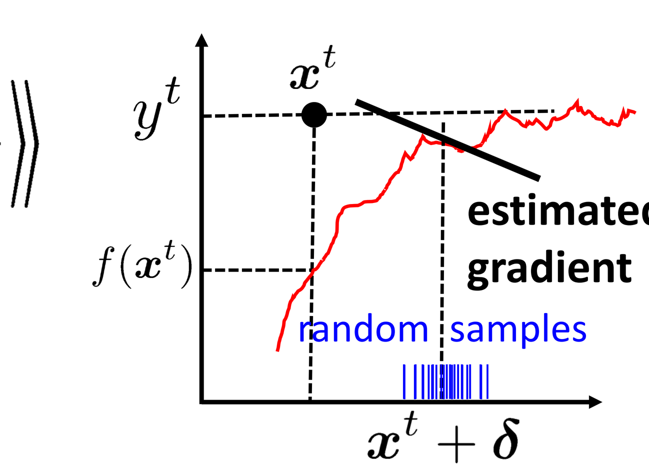
1. Local gradient estimation of f

- $f(\boldsymbol{x})$ is black-box but we need the gradient:

$$\frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\sigma_t^2} \left\langle\!\!\left\langle \frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \right\rangle\!\!\right\rangle + \lambda \boldsymbol{\delta}$$

- Gradient $\left\langle\!\left\langle \frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \right\rangle\!\right\rangle$

→ Local sampling & linear fit



2. Proximal gradient update for $\boldsymbol{\delta}$

- With the numerically estimated gradient, the problem now looks like L_1-regularized convex-*ish* optimization
- Updating rule from $\boldsymbol{\delta}^{\text{old}}$ using prox gradient-like algorithm

$$J(\boldsymbol{\delta}) \triangleq \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2\sigma_t^2} + \frac{1}{2}\lambda \|\boldsymbol{\delta}\|_2^2$$

$$\boldsymbol{\delta} = \arg\min_{\boldsymbol{\delta}} \left\{ J(\boldsymbol{\delta}^{\text{old}}) + (\boldsymbol{\delta} - \boldsymbol{\delta}^{\text{old}}) \langle\!\langle \nabla J(\boldsymbol{\delta}^{\text{old}}) \rangle\!\rangle + \frac{1}{2\kappa} \|\boldsymbol{\delta} - \boldsymbol{\delta}^{\text{old}}\|_2^2 + \nu \|\boldsymbol{\delta}\|_1 \right\}$$

$$= \text{prox}_{\kappa\nu\|\boldsymbol{\delta}\|_1} \left(\boldsymbol{\delta}^{\text{old}} - \kappa \langle\!\langle \nabla J(\boldsymbol{\delta}^{\text{old}}) \rangle\!\rangle \right)$$

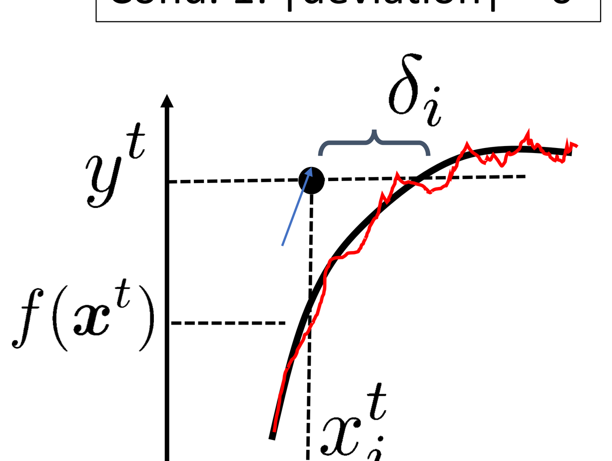
Has an analytic solution! (→ paper)

Condition of convergence – where the intuition of "horizontal deviation" comes from

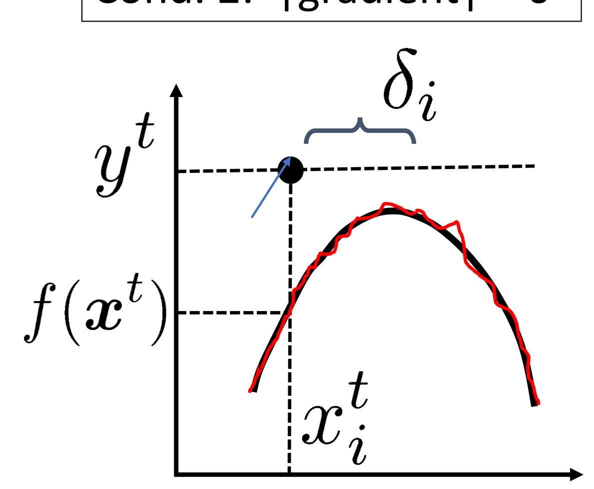
- The prox gradient-like update converges when

$$\frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\sigma_t^2} \left\langle\!\!\left\langle \frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \right\rangle\!\!\right\rangle \approx 0$$

Cond. 1: |deviation| = 0



Cond. 2: |gradient| = 0



Existing methods for anomaly attribution

- Few applicable approaches
- Existing methods are mostly for explaining f(x) in classification, often in white-box setting
- Possible baselines:

1: Z-score

$$a_i(\boldsymbol{x}^t) = \frac{x_i^t - \text{mean}_i}{\text{stddev}_i}$$

2. LIME [Ribeiro 18], extended to explain $f(\boldsymbol{x}^t) - y^t$ rather than $f(\boldsymbol{x}^t)$

3: Shapley value [Strumbelj+ 14], extended to explain $f(\boldsymbol{x}^t) - y^t$

Comparison in the building energy use-case

- One month-worth building energy data
  - y: energy consumption
  - x: time of day, temperature, humidity, sunrad, day of week (one-hot encoded)
- Top: Overall anomaly score
  - ≒ energy deviation
- Bottom 3: Responsibility score of each input variable
  - Only LC captures meaningful patterns